DOCUMENT RESUME

ED 409 354 TM 026 877

AUTHOR Seong, Tae-Je; And Others
TITLE A Comparison of Procedures for Ability Estimation under the Graded Response Model.
PUB DATE Mar 97
NOTE 44p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; Comparative Analysis; Difficulty Level; *Estimation (Mathematics); Foreign Countries; *Maximum Likelihood Statistics; Sample Size; Statistical Distributions; Test Length
IDENTIFIERS *Graded Response Model; A Posteriori Methods

ABSTRACT

This study was designed to compare the accuracy of three commonly used ability estimation procedures under the graded response model. The three methods, maximum likelihood (ML), expected a posteriori (EAP), and maximum a posteriori (MAP), were compared using a recovery study design for two sample sizes, two underlying ability distributions, and three test lengths. Recovery of ability was generally better for longer tests and for the conditions in which ability was matched to test difficulty. ML tended to recover less well than either EAP or MAP, particularly for the short test in the unmatched ability condition. For longer tests, all three methods recovered about equally well. (Contains 8 figures, 8 tables, and 26 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Tae-Je Seong
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 409 354

A Comparison of Procedures for Ability Estimation Under the Graded Response Model

Tae-Je Seong
Ewha Womans University, Seoul, Korea

Seock-Ho Kim
The University of Georgia

Allan S. Cohen
University of Wisconsin–Madison

March, 1997
Running Head: COMPARISON OF ABILITY ESTIMATION

Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

TM026877

BEST COPY AVAILABLE

A Comparison of Procedures for Ability Estimation Under the Graded Response Model

Abstract

This study was designed to compare the accuracy of three commonly used ability estimation procedures under the graded response model. The three methods, maximum likelihood(ML), expected a posteriori (EAP), and maximum a posteriori (MAP), were compared using a recovery study design for two sample sizes, two underlying ability distributions, and three test lengths. Recovery of ability was generally better for the longer tests and for the conditions in which ability was matched to test difficulty. ML tended to recover less well than either EAP or MAP, particularly for the short test in the unmatched ability condition. For longer tests, all three methods recovered about equally well.

Key words: ability estimation, graded response model, item response theory, MULTILOG.

Introduction

The majority of work on the evaluation and comparison of ability estimation procedures in item response theory (IRT) has been done using dichotomous models (e.g., Seong, 1990; Swaminathan & Gifford, 1983; Yen, 1987). (See Baker (1987) and Swaminathan (1983) for reviews of the estimation procedures for dichotomous IRT model.) Recent efforts to develop alternative measurement methods, such as performance assessment, however, have sparked interest in looking at other models. Several polytomous models have been proposed in the context of IRT. The graded response model (Samejima, 1969, 1972), the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the nominal response model (Bock, 1972) appear to offer some important promise for ability estimation in performance testing situations. Thissen and Steinberg (1986) and Dodd, De Ayala, and Koch (1995) present useful classifications of the different types of IRT models for polytomously scored items.

The majority of work on these models, however, has been done on item parameter estimation: Ankemann and Stone (1992) and Reise and Yu (1990) investigated parameter recovery for the graded response model; Choi, Cook, and Dodd (1996) for the partial credit model; and, De Ayala (1995) for the nominal response model. Results of these studies point to test length, sample size, and matched versus non-matched ability distributions as important factors in the accuracy of item parameter recovery. There do not appear to be any studies which have focused on accuracy of estimation procedures for ability parameters under the graded response model. In this study, therefore, we examined factors which affect ability estimation in the context of the graded response model.

Under Samejima's (1969, 1972) graded response model, the category response function $P_{jk}(\theta)$ is the probability of response k to item j as a function of θ. This function is defined as

$$P_{jk}(\theta) = \begin{cases} 1 - P^*_{j1}(\theta) & \text{when } k = 1 \\ P^*_{j(K-1)}(\theta) & \text{when } k = K \\ P^*_{j(k-1)}(\theta) - P^*_{jk}(\theta) & \text{otherwise,} \end{cases} \tag{1}$$

where $P^*_{jk}(\theta)$is the boundary response function in the form of the logistic model given by

$$P^*_{jk}(\theta) = \{1 + \exp[-\alpha_j(\theta - \beta_{jk})]\}^{-1}, \tag{2}$$

where α_j is the discrimination parameter for item j, β_{jk} is the location parameter, and θ is the trait level parameter. With $P^*_{j0}(\theta) = 1$ and $P^*_{jK} = 0$, the category response function can

be succinctly written as

$$P_{jk}(\theta) = P^{*}_{j(k-1)}(\theta) - P^{*}_{jk}(\theta), \tag{3}$$

where $k = 1(1)K$ and K is the total number of categories. Figures 1 and 2 illustrate the category response functions and the boundary response functions, respectively, for a typical graded response model item with five ordered response categories: $\alpha_j = 1.46$, $\beta_{j1} = -.35$, $\beta_{j2} = .67$, $\beta_{j3} = .97$, $\beta_{j4} = 1.94$.

Insert Figures 1 and 2 about here

Ability Estimation Under the Graded Response Model

The three ability estimation methods examined in this study were maximum likelihood (ML), expected a posterior (EAP), and maximum a posteriori (MAP).

EAP Estimation

The calculation of EAP estimates is relatively simple and noniterative (Bock & Aitkin, 1981; Bock & Mislevy, 1982). The EAP estimate is the mean of the posterior distribution of θ given either the vector of observed responses or a response pattern. Let y_j be the polytomous score for item j (i.e., $y_j = 1, 2, \ldots,$ or K) and let

$$u_{jk} = \begin{cases} 1 & \text{if } y_j = k \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

be the indicator variable for item j. Without loss of generality, we assume that all items in the test have the same number of categories K. The probability that $y_j = k$ at point θ on ability continuum is

$$\text{Prob}\{y_j = k|\theta\} = P_{jk}(\theta) = \prod_{k=1}^{K} P_{jk}(\theta)^{u_{jk}}. \tag{5}$$

The likelihood of θ given a response vector or a response pattern $y_l = (y_1, y_2, \ldots, y_J)$ can be written as

$$p(y_l|\theta) = L_l(\theta) = \prod_{j=1}^{J} \prod_{k=1}^{K} P_{jk}(\theta)^{u_{jk}}, \tag{6}$$

where J is the total number of items in the test.

The EAP estimate of θ with y_l can be approximated to any specified degree of precision with quadrature points X_q and weights $A(X_q)$ of a prior distribution using

$$\hat{\theta} = \frac{\sum_{q=1}^{Q} X_q L_l(X_q) A(X_q)}{\sum_{q=1}^{Q} L_l(X_q) A(X_q)}, \tag{7}$$

where Q is the total number of quadrature points. When a normal prior is assumed, Gauss-Hermite quadratures (Stroud & Secrest, 1966) or equally spaced quadratures and the corresponding weights can be applied. See also Bock and Mislevy (1982) for the posterior standard deviation formula.

ML Estimation

In maximum likelihood estimation, the likelihood function will be maximized to find an estimate of θ. Equivalently, we may work with the log likelihood function

$$L = \log p(y_l|\theta) = \sum_{j=1}^{J} \sum_{k=1}^{K} u_{jk} \log P_{jk}(\theta). \tag{8}$$

In order to obtain the θ estimate, we differentiate L with respect to θ and set it to zero. The equation cannot be solved directly as it is a nonlinear function of θ.

Newton's method can be used, however, to iteratively estimate θ:

$$\hat{\theta}_{(t)} = \hat{\theta}_{(t-1)} - \left[\frac{dL/d\theta}{d^2L/d\theta^2} \right]_{(t-1)}, \tag{9}$$

where (t) designates the iteration, and $dL/d\theta$ and $d^2L/d\theta^2$ are the first and second derivatives of L with respect to θ. (See Baker (1992a) for the required derivatives.)

MAP Estimation

If we assume a prior distribution of θ, we can obtain Bayes modal (MAP) estimates of ability. According to Bayes theorem,

$$p(\theta|y_l) \propto p(y_l|\theta) \times p(\theta), \tag{10}$$

where $p(\theta|y_l)$ is the posterior distribution, $\propto$ denotes proportionality, $p(y_l|\theta)$ is the likelihood function, and $p(\theta)$ is the prior distribution of θ. The posterior distribution is maximized to

obtain the MAP ability estimate. Equivalently, we can use the log posterior distribution and the posterior function F,

$$\log p(\theta|y_l) \propto \log p(y_l|\theta) + \log p(\theta) = F(\theta). \tag{11}$$

The Newton's equation is

$$\hat{\theta}_{(t)} = \hat{\theta}_{(t-1)} - \left[\frac{dF/d\theta}{d^2F/d\theta^2}\right]_{(t-1)}, \tag{12}$$

where (t) designates the iteration, and $dF/d\theta$ and $d^2F/d\theta^2$ are the first and second derivatives of F with respect to θ. For $p(\theta)$, the standard normal distribution was used in this study. That is,

$$p(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right). \tag{13}$$

The focus of this study was on accuracy of recovery of underlying ability parameters for three commonly used methods of ability estimation under the graded response model in the context of marginal maximum likelihood estimation (MMLE). In addition, item parameter recovery was investigated and compared with results from previous studies (e.g., Ankenmann & Stone, 1992; Reise & Yu, 1990).

Methods

Data Generation

Data for the simulation study were generated under the graded response model using the computer program GENIRV (Baker, 1988). The item parameters used to generate the data (see Table 1) were based on calibration results of the mathematics tests developed as a part of the Wisconsin Student Assessment System (Webb, 1994). All items had five ordered categories. The mean of the location parameters β_{jk} of the original 36 items was .962 and the standard deviation was .893.

Insert Table 1 about here

Three different lengths of tests were simulated: 5, 10, and 30 items. The items appearing on each of the three test lengths are identified in Table 1. Tests of all three lengths are common in different kinds of performance testing. Performance tests which require students

to show all of their work often have fewer items than tests which permit students to give shorter answers. A writing test, for example, which requires examinees to provide an outline and an essay that is written and then re-written or expanded one or more times in response to editorial comments, often can have as few as 5 items. Mathematics tests which require students to show only their answers can often include more items, sometimes as many as 30.

Items were selected so that the three tests all retained approximately the same level of difficulty. The means and standard deviations of the location parameters for each of the three lengths of tests are given at the bottom of Table 1. For the 5-item test, for example, the mean and standard deviation of the location parameters were .966 and .860, respectively. Note that the mean of the location parameters reflects the difficulty level of the test.

Data were generated for two sample sizes, 300 and 1,000 simulated examinees, with two underlying ability distributions: N(0,1) and N(1,1). For the purpose of this study, an approximation based on histograms was used to define the ability parameters instead of randomly sampling the underlying ability parameters from the specified normal distributions. There were 11 ability levels. Table 5 contains the number of examinees in each of the θ level for both 300 and 1,000 examinees with N(0,1) and N(1,1), respectively. The difficulty level of a well-designed test is typically matched to the ability of the examinee group. The N(1,1) distribution was essentially matched to the mean difficulty of the test. The test would be considered hard, however, for examinees whose ability distribution was N(0,1). 100 replications were generated for each sample size by ability distribution by test length ($2 \times 2 \times 3$) condition.

Item and Ability Parameter Estimation

Two of the most commonly used computer programs implementing the MMLE algorithm are MULTILOG (Thissen, 1991) and PARSCALE (Muraki & Bock, 1993). MULTILOG was used in this study for estimation of item and ability parameters. In MMLE (e.g., Bock & Aitkin, 1981), the likelihood is marginalized under the assumption that a population distribution exists. Default MULTILOG options under the graded response model were used for estimation of item parameters. MULTILOG provides marginal maximum likelihood item parameter estimates. The EAP estimates of ability can also be obtained in a single item and ability calibration run. Note that in the context of MMLE, all ability estimates (i.e., EAP, ML, and MAP) are obtained assuming item parameter estimates from the marginalized

likelihood are fixed and true values. Two additional MULTILOG runs for each data set were required to obtain ability estimates of ML and MAP. A standard normal prior was employed in ability estimation of EAP and MAP.

Linking Item and Ability Parameter Estimates

Since the IRT metric is inherently indeterminate, it was necessary to link parameter estimates onto a common metric before comparisons could be made. The test characteristic method for the graded response model (Baker, 1992b) as implemented in the computer program EQUATE (Baker, 1993) was used to link both item and ability parameter estimates to the metric of the underlying (i.e., generating) parameters. Once this was done, then ability parameter estimates could be compared among the three methods investigated in this study.

First, item and ability parameters were estimated and then linked to the underlying metric. Next, the computer program EQUATE was used to obtain the transformation coefficients A and B used to link the parameter estimates to the underlying metric. The transformation equations for the item parameters under the graded response model are

$$a_j^* = a_j / A \tag{14}$$

and

$$b_{jk}^* = A \times b_{jk} + B \tag{15}$$

where * indicates the transformed item parameter estimates on the underlying metric. Similarly, the transformation for the ability parameter estimate for person i ($i = 1, \ldots, N$) is

$$\hat{\theta}_i^* = A \times \hat{\theta}_i + B. \tag{16}$$

Since the test characteristic curve method of linking used only item parameters, only one EQUATE run had to be performed for the three ability estimation conditions. In case of the 5-item test, for example, the EQUATE run produced linking coefficients A and B based on the 5 items. Then, using the A and B coefficients from the EQUATE run, item parameter estimates for that replication as well as three sets of ability estimates, respectively, were transformed to the metric of the underlying parameters. A total of 1,200 EQUATE runs were performed, that is, 100 replications for the three test lengths of the N(0,1) and N(1,1) ability groups in each of the two combinations of sample sizes.

Accuracy of Item and Ability Parameter Estimates

In a recovery study such as this, it is possible to evaluate the quality of the results of different estimation methods by comparing the parameter estimates with the underlying parameters. Three indices were used in this study: Root mean square errors (RMSEs), statistical bias, and correlations between the parameter estimates and the underlying parameters.

RMSE for each item discrimination parameter, α_j, is defined as

$$\sqrt{\frac{1}{R}\sum_{r=1}^{R}(a_{jr}^{*}-\alpha_j)^2}, \tag{17}$$

where r designates the replication and $R = 100$ is the total number of replications used in this study.

Similarly, RMSE for an item location parameter, β_{jk}, is defined similarly as

$$\sqrt{\frac{1}{R}\sum_{r=1}^{R}(b_{jkr}^{*}-\beta_{jk})^2}. \tag{18}$$

RMSE for ability θ_i is defined as

$$\sqrt{\frac{1}{N_r}\sum_{r=1}^{N_r}(\hat{\theta}_{ir}^{*}-\theta_i)^2}, \tag{19}$$

where r designates the replication and N_r is the number of examinees assigned to θ_i for all 100 replications.

Bias is defined for an item discrimination parameter, α_j, as

$$\frac{1}{R}\sum_{r=1}^{R}(a_{jr}^{*}-\alpha_j), \tag{20}$$

where r designates the replication and $R = 100$ is the total number of replications.

Bias for an item location parameter, β_{jk}, is defined as

$$\frac{1}{R}\sum_{r=1}^{R}(b_{jkr}^{*}-\beta_{jk}). \tag{21}$$

Bias for ability θ_i is defined as

$$\frac{1}{N_r}\sum_{r=1}^{N_r}(\hat{\theta}_{ir}^{*}-\theta_i), \tag{22}$$

where r designates the replication and N_r is the number of examinees assigned to θ_i for all 100 replications.

Results

Recovery of Item Parameters

Average RMSEs for item parameters over the number of items in the test are given in Table 2. Figures 3 and 4 present RMSE results to illustrate the patterns of errors for item discrimination and item location, respectively. RMSEs for item discrimination were smaller for both the large sample and matched ability conditions. RMSEs were also smallest for the 30-item test, although differences were generally in the second and third decimal places. Likewise, RMSEs for location were smaller for the large sample and matched ability conditions. No differences were found for test length. In addition, RMSEs were smaller across all conditions for boundary location parameters that were relatively matched with the underlying ability parameters.

Insert Table 2 and Figures 3 and 4 About Here

The patterns of bias results for item discrimination and location parameters are shown in Figures 5 and 6, respectively. Average bias values for item parameters are given in Table 3. Biases for discrimination parameters were very small for all conditions simulated in the study. The large sample and the matched ability condition in general yielded better bias results for item discrimination. Bias values were also very small for location parameters. There was a tendency for bias results for location parameters to be slightly smaller in the large sample condition and in the matched ability condition. Essentially, MMLE yielded item parameter estimates with virtually no bias.

Insert Table 3 and Figures 5 and 6 About Here

Average correlations between generating parameters and parameter estimates over 100 replications are given in Table 4. In general, quality of recovery as indicated by correlations was similar to results for bias and RMSEs. Correlations for item discrimination were higher in the large sample and in the matched ability conditions. Correlations were smallest in the 5-item test. Recovery of location parameters tended to follow the same pattern as for discrimination except that correlations were very high for all test lengths. Correlations for the first three boundary location parameters were larger than correlations for the fourth.

Insert Table 4 About Here

These results indicate that recovery of item discrimination and location parameters was adequate. Recovery in the large sample and matched ability conditions was better, but recovery in the small sample and unmatched ability conditions was actually acceptable as well.

Recovery of Ability Parameters

In the context of ML estimation for 5-category graded response items, examinees whose responses are all 1's or 5's can not be estimated. The average numbers of such candidates over 100 replications are given in Tables 5a to 5d at the ability levels used to generate the data for each sample size for both matched and unmatched ability conditions.

The number of candidates generated at each ability level in the sample is indicated in the last line of each of these smaller tables. As an example, 20 candidates were generated at an ability of $-.5$ in the 300 examinee, matched ability condition (see Table 5b). An average of 2.44 of these examinees had non-finite ability estimates on the 5-item test. On the 10-item test, this average was .34 examinees and for the 30-item test, the average was 0. The results in Table 5 do not indicate the quality of recovery of ability parameters under ML estimation but they do indicate where the bulk of the data were located and how much data were recovered by the ML method.

Insert Table 5 About Here

The pattern of RMSEs across ability can be seen in Figure 7. For the 10-item test, RMSEs were larger at the ends of the ability distribution and smaller in the middle of the distribution. This same general pattern was found for both the 5-item and 30-item tests.

Insert Figure 7 About Here

Average RMSEs in ability estimates are given in Table 6. Recovery of underlying ability parameters did not appear to differ by methods of estimation. There was a clear effect on recovery, however, for test length: RMSEs decreased with an increase in test length. In

addition, RMSEs were slightly smaller for the matched ability condition. Sample size did not appear to have an effect on magnitude of the RMSEs in Table 6.

Insert Table 6 About Here

The pattern of bias statistics across ability levels can be seen clearly in Figure 8. For example, for the 5-item test in the small sample condition, recovery was poorer for the low ability examinees in the unmatched ability condition. This same general pattern was observed for the 10-item and 30-item test lengths as well. Sample size did not appear to have an effect on recovery of ability parameters. Bias values were smaller, however, for all methods in the matched ability condition and decreased as test length increased. ML tended to yield relatively smaller sizes of bias than did EAP and MAP except 5-item test with N(0,1). Bias patterns from EAP and MAP are very similar. EAP in general yielded smaller bias than did MAP. Note that there are many cases where ML estimation was not possible.

Insert Figure 8 About Here

The average biases for ability parameter estimates for each of the three estimation methods are given in Table 7. Bias statistics decreased as test length increased. Sample size did not appear to have an effect on recovery of ability parameters. Bias statistics were also smaller for all methods in the matched ability condition and for the EAP method in the unmatched ability condition in both large and small samples.

Insert Table 7 About Here

Average correlations between generating parameters and ability estimates over 100 replications are given in Table 8. The average number of non-finite cases excluded under ML is also given in the right column in Table 8. As an example, an average of 44.34 cases were excluded from each replication for the 5-item test in the small sample unmatched ability condition. In general, the results appear to be consistent across estimation method. ML estimation did have slightly lower correlations in the 5-item length test due to exclusion of examinees but these differences were quite small and are essentially negligible. Recovery as

measured by correlations improved with an increase in test length. Both test length and ability matching condition appeared to have some impact on recovery.

Insert Table 8 About Here

Discussion

The comparability of ability estimates under IRT across different estimation algorithms is an important concern for test developers. The expectation is that ability estimates should be the same across all methods. Whether or not this is the case for the graded response model, however, is not clear. The only work that has been reported on this subject has been done in the context of dichotomous models. This is unfortunate given the importance of polytomous models in alternative measurement procedures such as performance assessment. Several different methods are available for ability estimation for the graded response model. In this paper, we have compared simulation results from three of the more commonly used methods, maximum likelihood estimation, expected a posteriori estimation and maximum a posteriori estimation.

The recovery study approach used here permitted comparison to be made between the generating parameters and the estimates of those parameters. The simulation results indicated that recovery of item parameters was good. Consistent with previous research, test length was an important factor in recovery of the discrimination parameters. Recovery of discrimination parameters was not as good for the short, 5-item test as for the 10- or 30-item tests. Recovery of both discrimination and location parameters was better in the large sample size and the matched ability conditions.

Recovery of ability parameters was generally better in the longer tests. Sample size appeared to have had little effect on recovery. Test length, however, did have an effect: Recovery for the 10- and 30-item tests was better. Few differences were found among the three methods in recovery of ability parameters. As might be expected, ML was not able to estimate ability for examinees with answers that were all 1s or all 5s. In the unmatched ability condition, bias statistics suggested that ML estimation recovered less well for the short, 5-item test than either EAP or MAP. But, both ML and MAP recovered underlying ability parameters better for the 10- and 30-item tests in the unmatched ability conditions. These results were not present, however, in either the RMSEs or correlations.

Results from the present suggest that, in general, ability estimation using any of the three algorithms is appropriate when ability is well-matched to test difficulty. When this is not the case, based on the bias results, choice of algorithm may need to take test length into account. That is, when examinees are tested with a hard test, ability estimation for short tests might be better with either the EAP or MAP methods. Results of this study are encouraging in that recovery under all simulated conditions was generally quite good. That is, the ability estimation algorithms implemented in MULTILOG appeared to be relatively robust under the conditions simulated here.

References

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded response model.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, *11*, 111–141.

Baker, F. B. (1988). *GENIRV: A program to generate item response vectors* [Computer program]. Madison, University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design.

Baker, F. B. (1992a). *Item response theory: Parameter estimation techniques.* New York: Dekker.

Baker, F. B. (1992b). Equating tests under the graded response model. *Applied Psychological Measurement*, *16*, 87–96.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement*, *17*, 20.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., & Mislevy, R. M. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Choi, S. W., Cook, K. F., & Dodd, B. G. (1996, April). *Parameter recovery for the partial credit model using MULTILOG.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

De Ayala, R. J. (1995, April). *Item parameter recovery for the nominal response model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19,* 5–22.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks* [Computer program]. Chicago: Scientific Software.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement,* No. 18.

Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distribution. *Applied Psychological Measurement, 14,* 299–311.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliff, NJ: Prentice-Hall.

Swaminathan, H. (1983). Parameter estimation in item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 24–44). Vancouver, Canada: Educational Research Institute of British Columbia.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). New York: Academic Press.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software.

Thissen, D., & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika, 51,* 567–577.

Webb, N. L. (1994). *Wisconsin performance assessment development project: Analysis and technical report for fiscal year 1993-94*. Madison: University of Wisconsin, Wisconsin Center for Educational Research.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, *52*, 275–291.

Figure Captions

Figure 1. Category Response Functions for a Five-Category Item

Figure 2. Boundary Response Functions for a Five-Category Item

Figure 3. Pattern of RMSEs for Item Discrimination Parameter Estimates

Figure 4. Pattern of RMSEs for Item Location Parameter Estimates

Figure 5. Pattern of Bias Results for Item Discrimination Parameter Estimates

Figure 6. Pattern of Bias Results for Item Location Parameter Estimates

Figure 7. Pattern of RMSEs for Ability Parameter Estimates

Figure 8. Pattern of Biases for Ability Parameter Estimates

Figure 1. Category Response Functions for a Five-Category Item

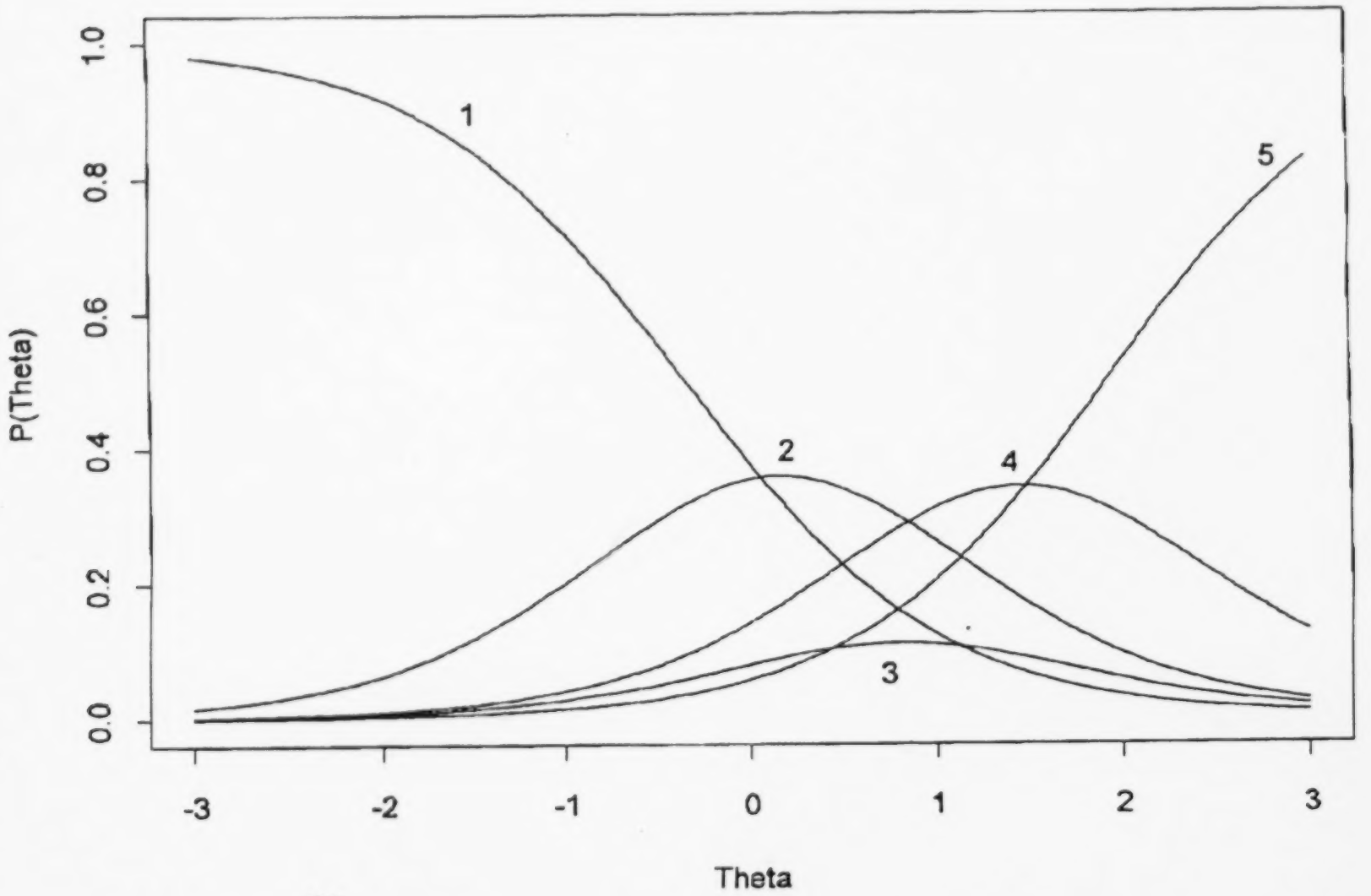
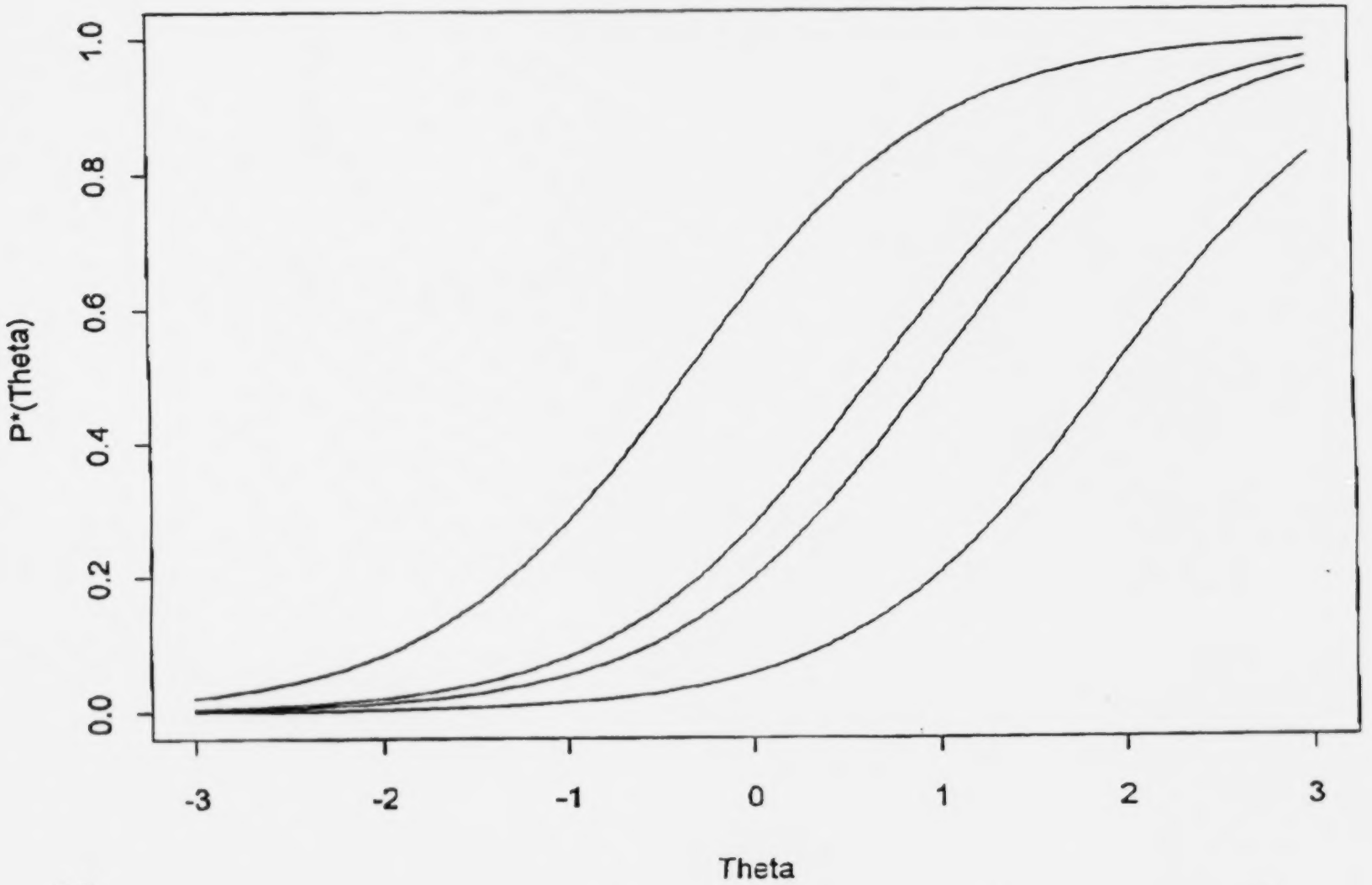


Figure 2. Boundary Response Functions for a Five-Category Item



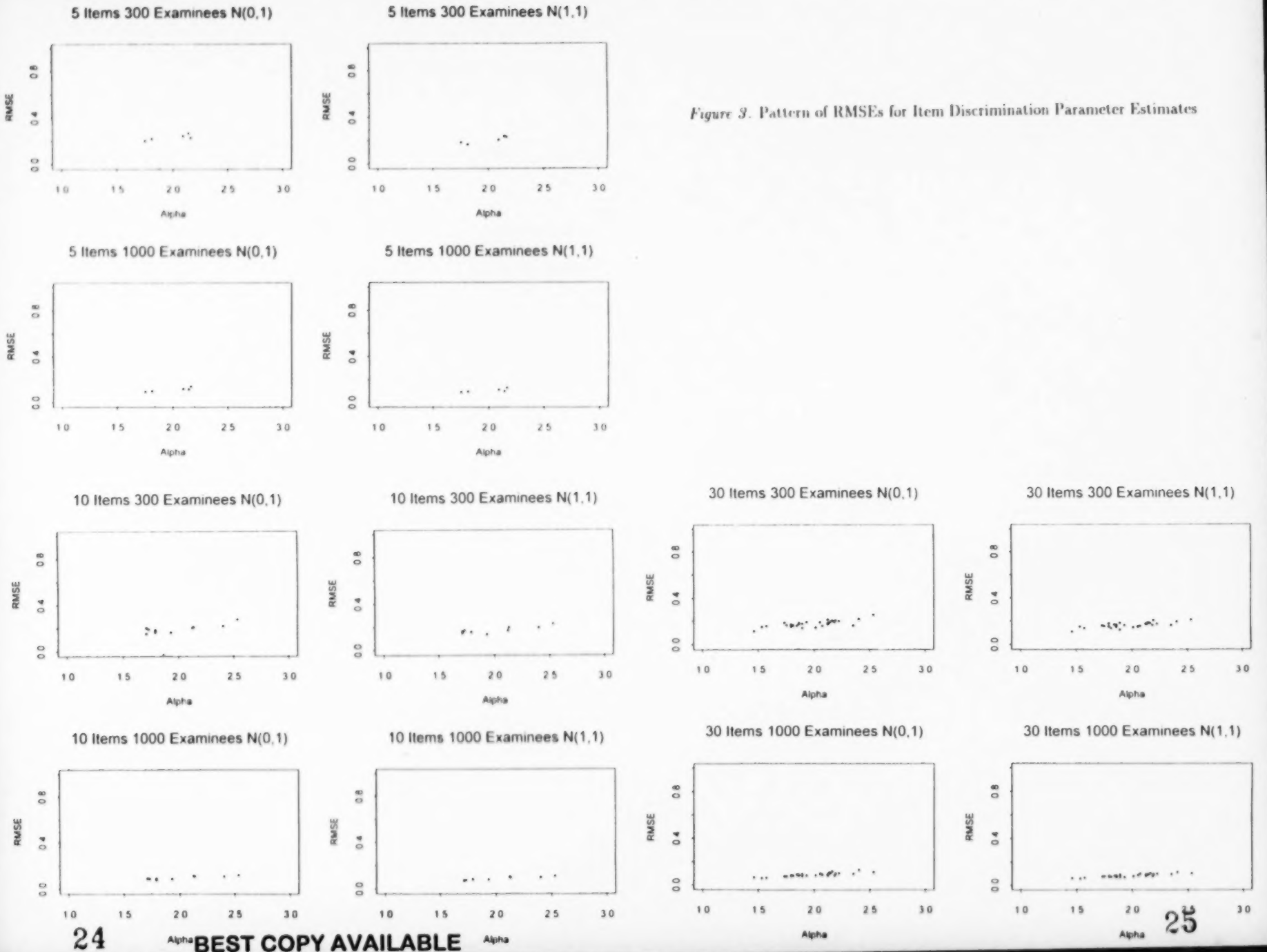

5 Items 300 Examinees N(0,1)
5 Items 300 Examinees N(1,1)
5 Items 1000 Examinees N(0,1)
5 Items 1000 Examinees N(1,1)
10 Items 300 Examinees N(0,1)
10 Items 300 Examinees N(1,1)
30 Items 300 Examinees N(0,1)
30 Items 300 Examinees N(1,1)
10 Items 1000 Examinees N(0,1)
10 Items 1000 Examinees N(1,1)
30 Items 1000 Examinees N(0,1)
30 Items 1000 Examinees N(1,1)
RMSE
Alpha

Figure 3. Pattern of RMSEs for Item Discrimination Parameter Estimates

BEST COPY AVAILABLE

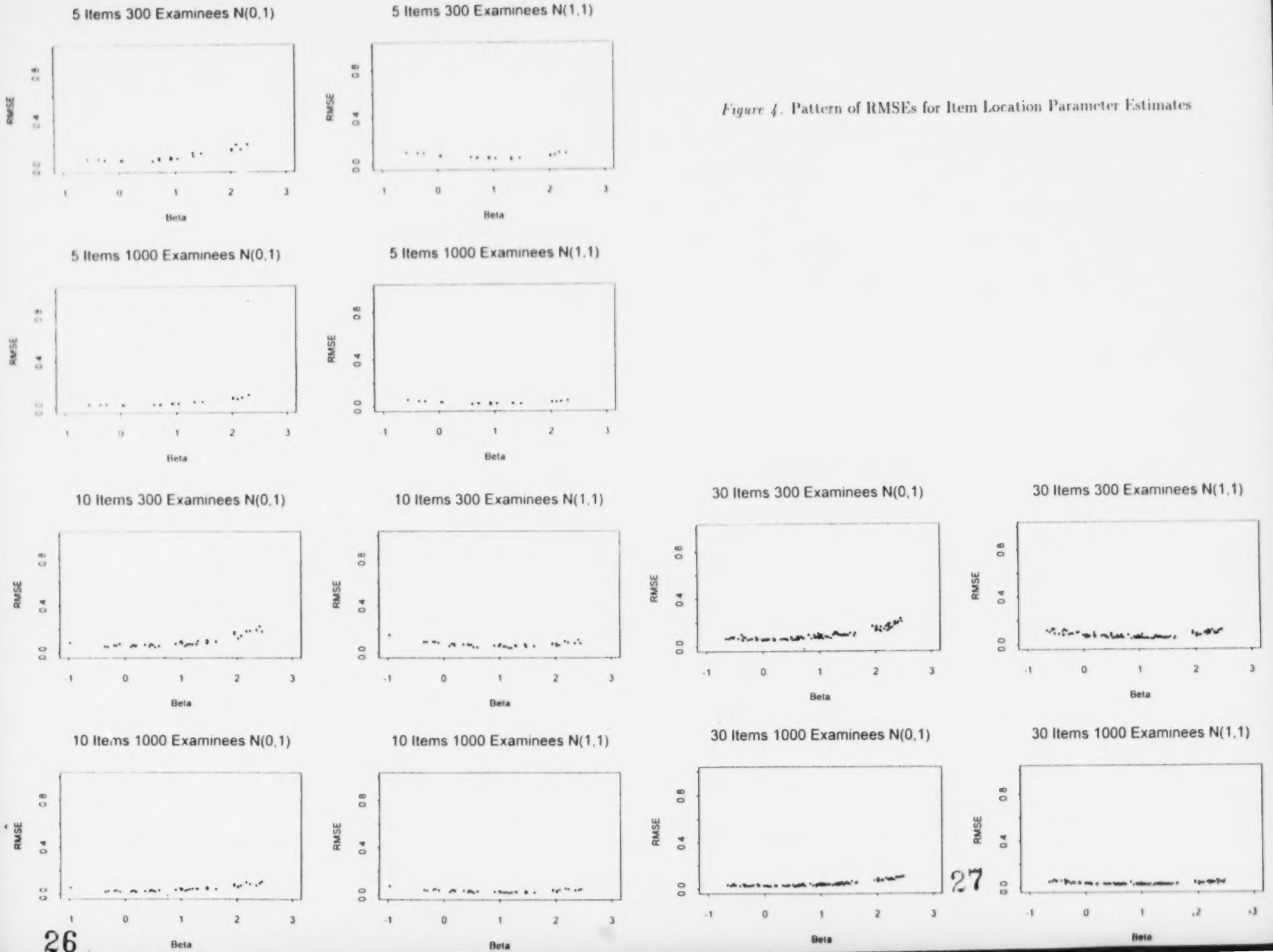

5 Items 300 Examinees N(0,1)
5 Items 300 Examinees N(1,1)
5 Items 1000 Examinees N(0,1)
5 Items 1000 Examinees N(1,1)
10 Items 300 Examinees N(0,1)
10 Items 300 Examinees N(1,1)
30 Items 300 Examinees N(0,1)
30 Items 300 Examinees N(1,1)
10 Items 1000 Examinees N(0,1)
10 Items 1000 Examinees N(1,1)
30 Items 1000 Examinees N(0,1)
30 Items 1000 Examinees N(1,1)
RMSE
Beta

Figure 4. Pattern of RMSEs for Item Location Parameter Estimates

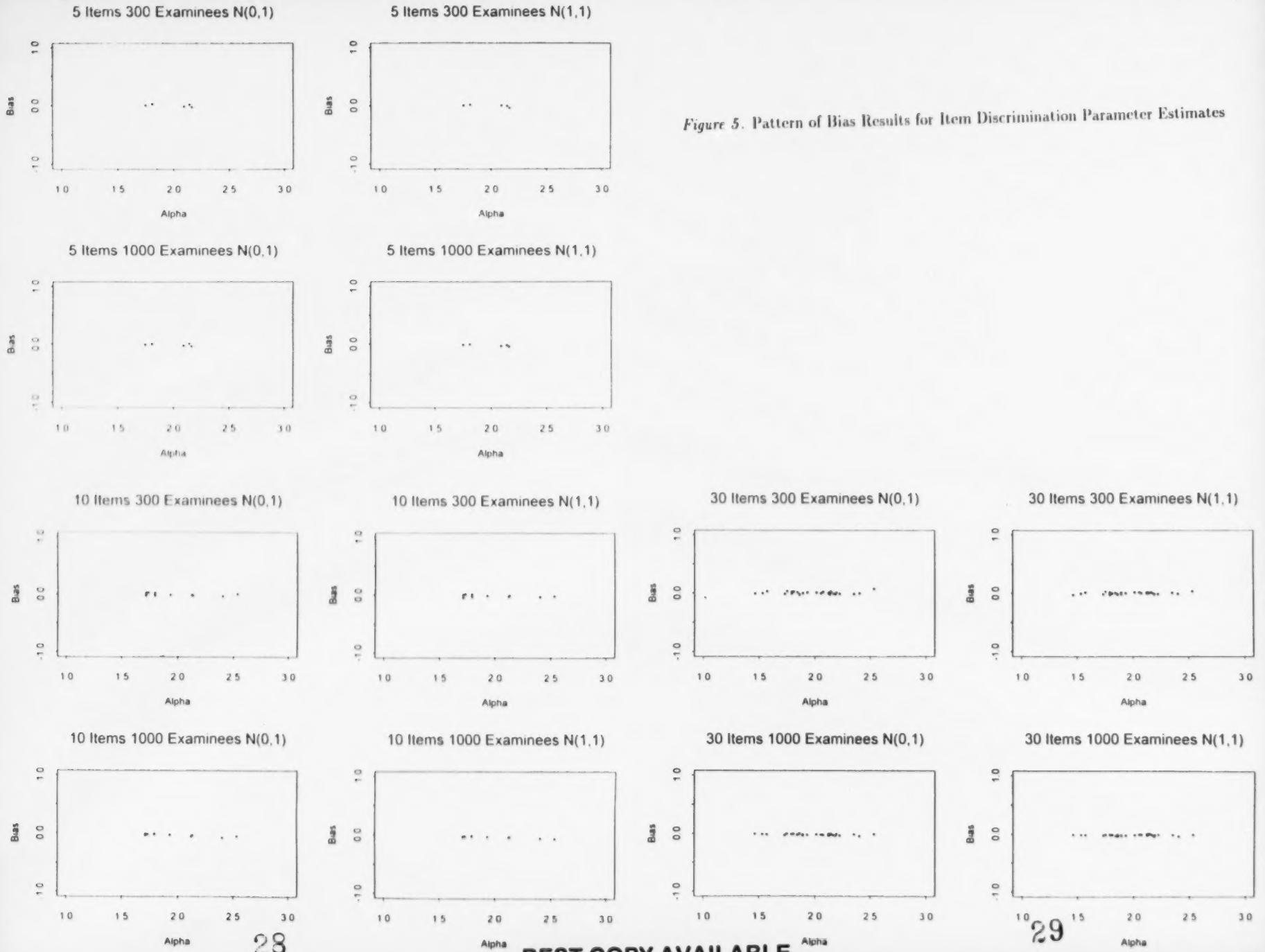
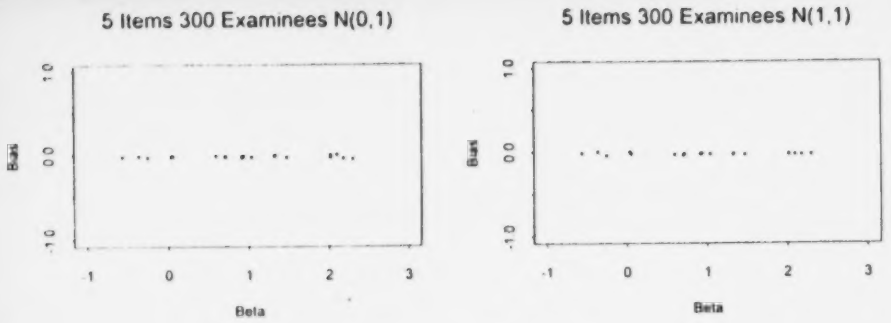
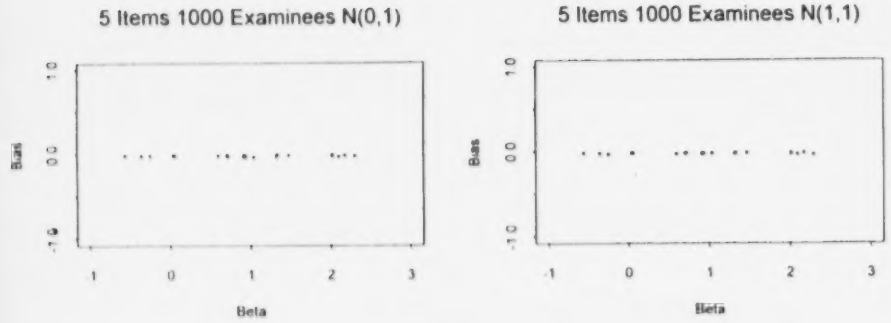
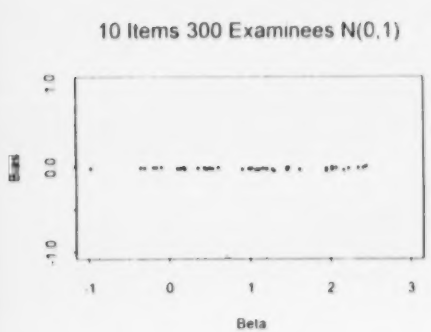


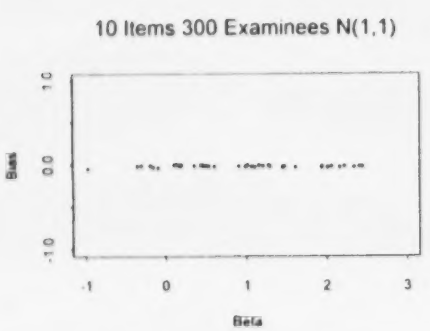
Figure 5. Pattern of Bias Results for Item Discrimination Parameter Estimates

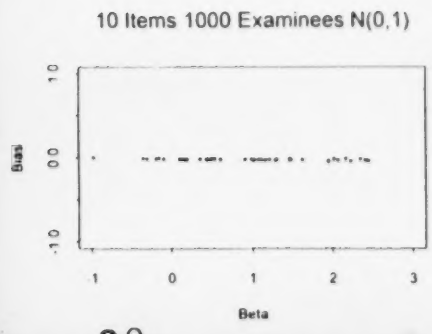
BEST COPY AVAILABLE











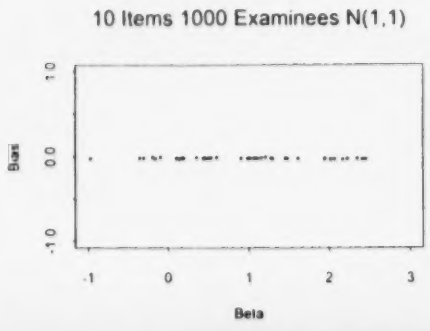
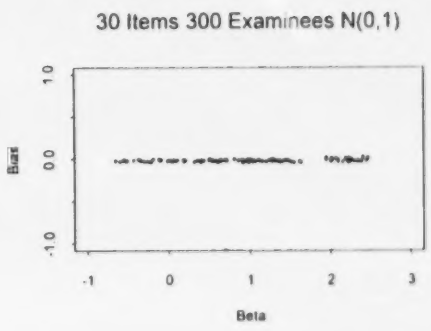
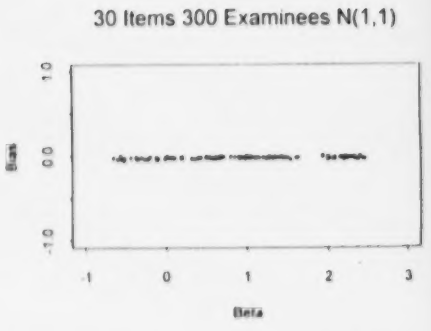
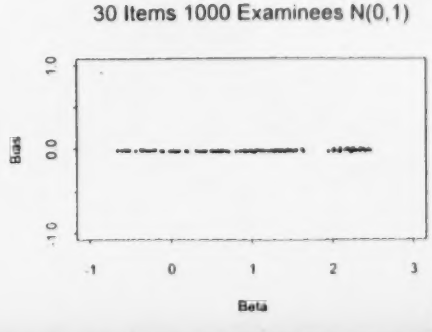
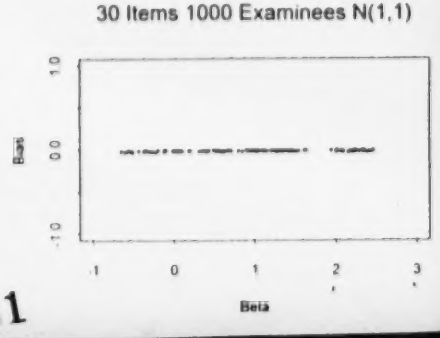


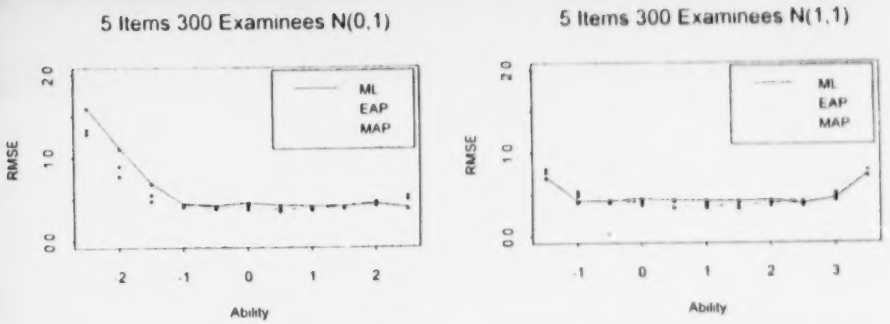
Figure 6. Pattern of Bias Results for Item Location Parameter Estimates

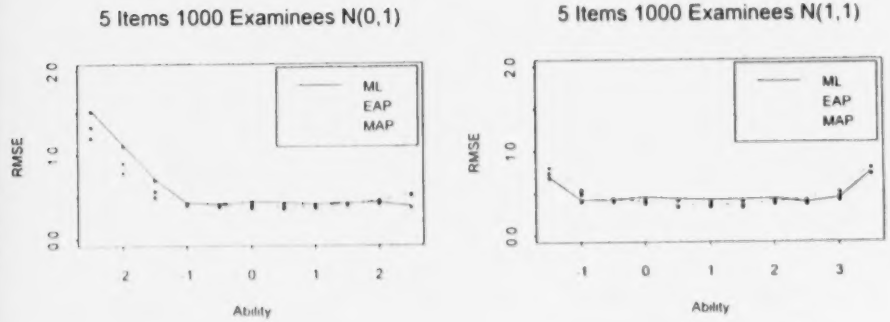


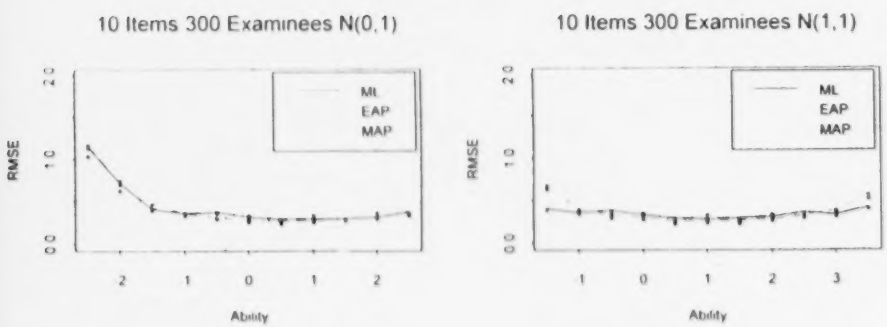












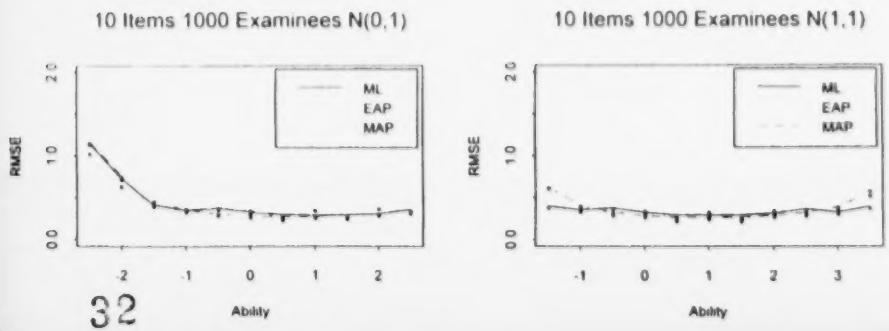
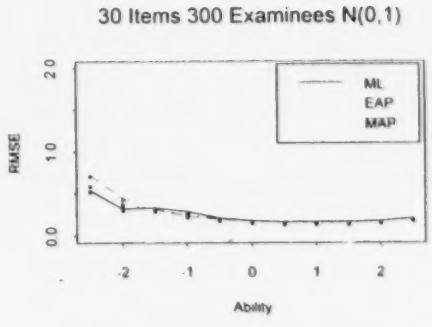
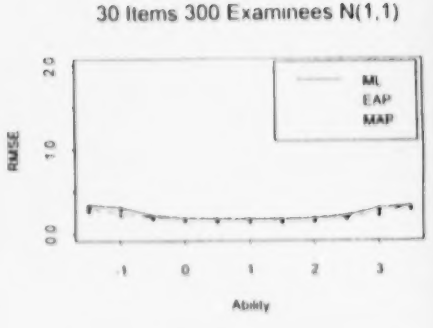
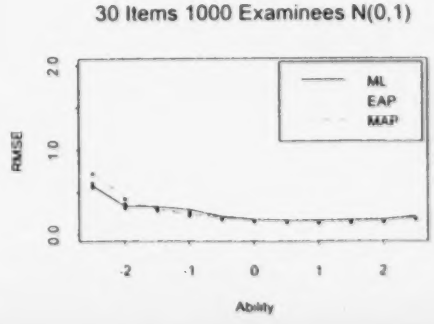
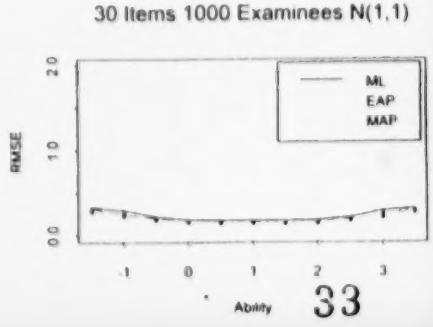


Figure 7. Pattern of RMSEs for Ability Estimates









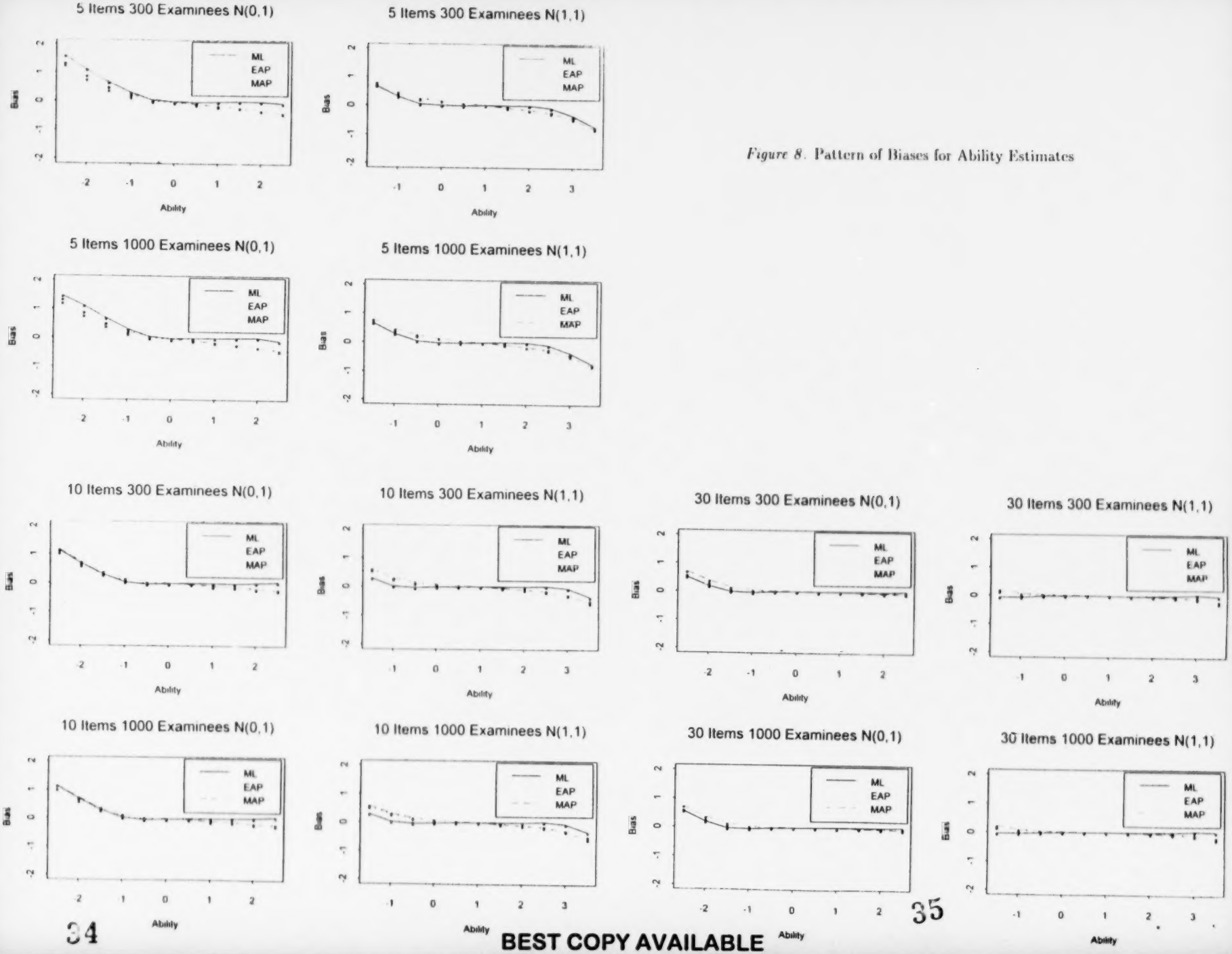


Figure 8. Pattern of Biases for Ability Estimates

BEST COPY AVAILABLE

Table 1

Item Parameters Estimates for Spring, 1994 Mathematics Field Test and Generating Item Parameters for 5-, 10- and 30-Item Tests

Item No.	Parameter					Test Length		
	α_j	β_{j1}	β_{j2}	β_{j3}	β_{j4}	5-Item	10-Item	30-Item
1	1.46	−.35	.67	.97	1.94			1
2	1.73	.18	.90	1.29	1.94		1	2
3	1.81	−.37	.03	.91	2.29	1		3
4	1.53	−.56	−.13	.80	2.22			4
5	1.57	−.38	.49	1.04	2.33			5
6	1.89	−.61	.63	1.37	2.34			6
7	1.89	.01	.67	1.33	2.18			7
8	1.84	−.23	.31	.98	2.46			8
9	1.71	−.98	−.16	1.45	1.94		2	
10	1.93	−.31	.60	1.27	2.44		3	9
11	2.53	−.36	.53	1.20	2.34		4	10
12	2.09	−.26	.70	1.46	2.17	2		
13	2.42	.24	.70	1.41	2.14			
14	1.79	−.52	.39	1.54	2.00			11
15	1.86	−.53	−.12	1.27	2.25			12
16	2.35	.06	.99	1.50	2.20			13
17	1.79	−.20	.49	1.00	2.40		5	14
18	2.12	.20	.56	1.40	2.00			15
19	2.07	−.44	.18	1.34	2.15			16
20	1.95	−.04	.74	1.14	2.30			
21	2.19	−.01	.39	1.36	2.01			17
22	2.40	.10	1.06	1.61	2.01		6	18
23	1.79	−.10	.35	1.01	2.22		7	19
24	2.12	.19	1.10	1.45	2.01		8	20
25	1.75	−.57	.93	1.31	2.01	3		21
26	2.16	.59	.91	1.32	2.01	4		22
27	1.86	−.02	.63	1.28	2.01			23
28	1.71	.14	.45	.98	2.16		9	
29	2.22	.52	.85	1.43	2.01			24
30	2.18	−.27	.58	1.24	2.25			25
31	2.01	−.66	.41	1.63	2.24			26
32	2.14	.05	.71	1.03	2.09	5		27
33	2.13	.43	1.15	1.47	2.06		10	28
34	2.12	.08	.70	1.12	2.09			29
35	1.95	−.18	.78	1.31	2.01			
36	2.05	.19	.61	.94	2.38			30
Mean	1.975	−.138	.577	1.254	2.156	.966	.995	.996
SD	.252	.355	.330	.215	.154	.860	.879	.887

BEST COPY AVAILABLE

Table 2
Average Root Mean Square Errors
In Item Parameter Estimation

Examinee	Ability	Item	Parameter				
			α_j	β_{j1}	β_{j2}	β_{j3}	β_{j4}
300	N(0,1)	5	.244	.075	.079	.109	.176
		10	.209	.077	.080	.105	.184
		30	.190	.077	.082	.109	.178
	N(1,1)	5	.216	.103	.075	.069	.101
		10	.185	.101	.075	.074	.095
		30	.176	.103	.080	.075	.106
1,000	N(0,1)	5	.126	.042	.043	.058	.099
		10	.110	.041	.047	.061	.098
		30	.100	.041	.044	.058	.096
	N(1,1)	5	.116	.054	.038	.038	.053
		10	.100	.055	.043	.039	.058
		30	.092	.057	.042	.041	.056

Table 3

Average Bias in Item Parameter Estimation

Examinee	Ability	Item	Parameter				
			α_j	β_{j1}	β_{j2}	β_{j3}	β_{j4}
300	N(0,1)	5	.027	.022	.004	.003	.001
		10	.021	.006	.002	-.001	-.006
		30	.029	.003	.002	-.001	.008
	N(1,1)	5	.018	.003	.003	-.000	-.002
		10	.011	.002	-.000	.000	-.004
		30	.024	-.004	-.000	.001	.001
1,000	N(0,1)	5	.007	.003	.002	-.002	.001
		10	-.006	.008	.003	-.002	-.003
		30	.009	.001	.000	-.001	.003
	N(1,1)	5	.002	.001	.001	.002	-.004
		10	-.009	.005	.003	-.000	-.001
		30	.006	-.001	-.000	.000	.000

Table 4

Average Correlations Between Item Parameters and Estimates

Examinee	Ability	Item	Parameter				
			α_j	β_{j1}	β_{j2}	β_{j3}	β_{j4}
300	N(0,1)	5	.559	.988	.978	.909	.449
		10	.810	.984	.982	.906	.734
		30	.803	.976	.970	.892	.662
	N(1,1)	5	.601	.975	.982	.960	.750
		10	.848	.973	.984	.948	.896
		30	.828	.957	.972	.946	.826
1,000	N(0,1)	5	.826	.997	.994	.972	.738
		10	.932	.995	.994	.965	.889
		30	.927	.993	.991	.968	.859
	N(1,1)	5	.841	.992	.996	.988	.911
		10	.944	.992	.995	.986	.958
		30	.939	.986	.992	.983	.943

Table 5

Average Number of Non-Finite Ability Estimates from Maximum Likelihood Estimation in 100 Replications

Table 5a: 300 Examinees N(0,1)

	Underlying Ability										
No. of Items	−2.5	−2.0	−1.5	−1.0	−.5	0	.5	1.0	1.5	2.0	2.5
5	3.70	6.75	13.46	13.20	5.59	.85	0	0	.01	.17	.61
10	3.41	5.69	8.30	4.94	.73	.01	0	0	0	0	.03
30	2.65	2.54	1.75	.10	0	0	0	0	0	0	0
No. of Examinees	4	8	20	36	52	60	52	36	20	8	4

Table 5b: 300 Examinees N(1,1)

	Underlying Ability										
No. of Items	−1.5	−1.0	−.5	0	.5	1.0	1.5	2.0	2.5	3.0	3.5
5	2.52	2.97	2.44	.51	.02	0	.05	.59	2.80	3.60	2.89
10	1.43	1.18	.34	0	0	0	0	.01	.36	1.36	2.07
30	.27	.02	0	0	0	0	0	0	0	.01	.37
No. of Examinees	4	8	20	36	52	60	52	36	20	8	4

Table 5c: 1000 Examinees N(0,1)

	Underlying Ability										
No. of Items	−2.5	−2.0	−1.5	−1.0	−.5	0	.5	1.0	1.5	2.0	2.5
5	11.30	23.66	43.54	45.44	19.62	2.78	.13	0	.03	.55	1.62
10	10.40	19.41	28.29	17.24	2.54	.04	0	0	0	0	.17
30	7.60	10.29	5.15	.33	0	0	0	0	0	0	0
No. of Examinees	12	28	66	121	174	198	174	121	66	28	12

Table 5d: 1000 Examinees N(1,1)

	Underlying Ability										
No. of Items	−1.5	−1.0	−.5	0	.5	1.0	1.5	2.0	2.5	3.0	3.5
5	7.75	10.51	6.99	1.55	.11	0	.07	2.26	9.32	12.58	8.71
10	4.84	3.81	.94	.03	0	0	0	0	1.18	4.64	5.84
30	.92	.06	0	0	0	0	0	0	0	.06	1.19
No. of Examinees	12	28	66	121	174	198	174	121	66	28	12

Table 6

Average Root Mean Square Errors in Ability Estimates

Examinee	Ability	Item	Method			
			EAP	MAP	ML	$\pm\infty$[a]
300	N(0,1)	5	.450	.441	.446	44.34
		10	.352	.335	.334	23.11
		30	.206	.208	.209	7.04
	N(1,1)	5	.410	.402	.430	18.39
		10	.322	.298	.315	6.75
		30	.174	.179	.188	.67
1,000	N(0,1)	5	.444	.438	.442	148.67
		10	.349	.333	.334	78.09
		30	.205	.206	.210	23.37
	N(1,1)	5	.405	.399	.429	59.85
		10	.316	.296	.314	21.28
		30	.171	.178	.186	2.23

[a]Average number of non-finite ability estimates under ML.

Table 7

Average Bias in Ability Parameter Estimates

Examinee	Ability	Item	Method			
			EAP	MAP	ML	$\pm\infty$[a]
300	N(0,1)	5	.005	.050	.130	44.34
		10	.024	.059	.051	23.11
		30	.014	.036	.002	7.04
	N(1,1)	5	-.001	.000	-.007	18.39
		10	.004	.006	.001	6.75
		30	-.001	-.001	-.001	.67
1,000	N(0,1)	5	.005	.049	.130	148.67
		10	.019	.051	.049	78.09
		30	.012	.035	.002	23.37
	N(1,1)	5	.002	.003	-.006	59.85
		10	.000	.001	-.002	21.28
		30	.000	.000	.000	2.23

[a]Average number of non-finite ability estimates under ML.

Table 8

Average Correlations Between Ability Parameters and Estimates

Examinee	Ability	Item	Method			
			EAP	MAP	ML	$\pm\infty$[a]
300	N(0,1)	5	.887	.891	.867	44.34
		10	.932	.938	.932	23.11
		30	.976	.976	.975	7.04
	N(1,1)	5	.911	.914	.894	18.39
		10	.945	.954	.949	6.75
		30	.985	.984	.983	.67
1,000	N(0,1)	5	.889	.891	.869	148.67
		10	.932	.938	.932	78.09
		30	.976	.976	.975	23.37
	N(1,1)	5	.912	.914	.894	59.85
		10	.946	.954	.949	21.28
		30	.985	.984	.983	2.23

[a]Average number of cases excluded under ML.

Acknowledgments

The authors express their gratitude to Frank B. Baker for making his EQUATE and GENIRV programs available for this study.

Authors' Addresses

Send requests for reprints or further information to Tae-Je Seong, Ewha Womans University, Seoul, Korea, Seock-Ho Kim, The University of Georgia, 325 Aderhold Hall, Athens, GA 30602, or Allan S. Cohen, Testing and Evaluation Services, University of Wisconsin, 1025 West Johnson Street, Madison, WI 53706. Internet: seong@tne.edsci.wisc.edu, skim@coe.uga.edu, or cohen@tne.edsci.wisc.edu